177

Reihe Ökonomie

Economics Series

Approaches for the Joint Evaluation of Hypothesis Tests:

Classical Testing, Bayes Testing, and Joint Confirmation

Robert M. Kunst

177

Reihe Ökonomie

Economics Series

Approaches for the Joint Evaluation of Hypothesis Tests:

Classical Testing, Bayes Testing, and Joint Confirmation

Robert M. Kunst

September 2005

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Contact:

Robert M. Kunst
Department of Economics and Finance
Institute for Advanced Studies
Stumpergasse 56
1060 Vienna, Austria
and
University of Vienna
Department of Economics
Brünner Straße 72
1210 Vienna, Austria
email: robert.kunst@univie.ac.at

Founded in 1963 by two prominent Austrians living in exile – the sociologist Paul F. Lazarsfeld and the economist Oskar Morgenstern – with the financial support from the Ford Foundation, the Austrian Federal Ministry of Education and the City of Vienna, the Institute for Advanced Studies (IHS) is the first institution for postgraduate education and research in economics and the social sciences in Austria. The **Economics Series** presents research done at the Department of Economics and Finance and aims to share "work in progress" in a timely way before formal publication. As usual, authors bear full responsibility for the content of their contributions.

Das Institut für Höhere Studien (IHS) wurde im Jahr 1963 von zwei prominenten Exilösterreichern – dem Soziologen Paul F. Lazarsfeld und dem Ökonomen Oskar Morgenstern – mit Hilfe der Ford-Stiftung, des Österreichischen Bundesministeriums für Unterricht und der Stadt Wien gegründet und ist somit die erste nachuniversitäre Lehr- und Forschungsstätte für die Sozial- und Wirtschafts-wissenschaften in Österreich. Die **Reihe Ökonomie** bietet Einblick in die Forschungsarbeit der Abteilung für Ökonomie und Finanzwirtschaft und verfolgt das Ziel, abteilungsinterne Diskussionsbeiträge einer breiteren fachinternen Öffentlichkeit zugänglich zu machen. Die inhaltliche Verantwortung für die veröffentlichten Beiträge liegt bei den Autoren und Autorinnen.

Abstract

The occurrence of decision problems with changing roles of null and alternative hypotheses has increased interest in extending the classical hypothesis testing setup. Particularly, confirmation analysis has been in the focus of some recent contributions in econometrics. We emphasize that confirmation analysis is grounded in classical testing and should be contrasted with the Bayesian approach. Differences across the three approaches – traditional classical testing, Bayes testing, joint confirmation – are highlighted for a popular testing problem. A decision is searched for the existence of a unit root in a time-series process on the basis of two tests. One of them has the existence of a unit root as its null hypothesis and its non-existence as its alternative, while the roles of null and alternative are reversed for the other hypothesis test.

Keywords

Confirmation analysis, decision contours, unit roots

JEL Classifications

C11, C12, C22, C44.

Contents

1 Introduction

The occurrence of decision problems with role reversal of null and alternative hypotheses has increased the interest in extensions of the classical hypothesis testing setup. Particularly, confirmation analysis has been in the focus of some recent econometric works (see Dhrymes, 1998, Keblowski and Welfe, 2004, among others). This paper analyzes the contribution of confirmation analysis against the background of the more general statistical decision problem, and compares it to alternative solution concepts. The focus is on the decision for a unit root in time-series processes.

In this and in comparable situations, the basic difficulty of traditional hypothesis testing is that pairs of tests may lead to a contradiction in their individual results. The role reversal of null and alternative hypothesis prevents the traditional construction of a new test out of the two components that may dominate individual tests with regard to power properties. The existing literature appears to suggest that conflicting outcomes indicate possible invalidity of the maintained hypothesis and therefore imply 'no decision' (e.g., see Hatanaka, 1996). In empirical applications, the *ad hoc* decision on the basis of comparing p–values is also widespread.

In a simplified interpretation, confirmation analysis (Charemza and Syczewska, 1998, Keblowski and Welfe, 2004, Dhrymes, 1998) suggests to select one of the two hypotheses as the overall null, thereby apparently resolving the conflict. An obvious difficulty of this approach is that a 'generic' alternative of one of the component tests is condensed to a lower-dimensional null by choosing specific parts of the alternative. This selection may seem artificial. A logical drawback is also that the confirmation test follows the classical asymmetry of one of the component tests, while the basic problem reveals a symmetric construction principle. If the basic problem clearly indicated the choice of null and alternative, the application of a test with role reversal would not be adequate at all.

A different and completely symmetric solution is Bayes testing. It is well known from the literature on statistical decision theory (Lehmann and Romano, 2005, Ferguson, 1967, Pratt *et al.*, 1995) that Bayes tests define a complete class for any given loss function, in the sense that any other test is dominated by a Bayes test. However, Bayes tests require the specification of elements such as loss functions and prior distributions that are often regarded as subjective.

This paper adopts a Bayesian viewpoint and presents Bayes-test solutions to the unit-root decision problem in a graphical manner in rectangles of null fractiles. In a related decision problem on the existence of seasonal unit roots, a similar approach was introduced by Kunst and Reutter (2002). Using some sensitivity checks by varying prior distributions, the relative benefits with regard to loss criteria can be compared to the classical and also to the confirmation approach. The graphical representations allow a convenient simplification and

visualization of traditional Bayes tests by focusing exclusively on the observed test statistics.

Section 2 analyzes the three approaches to joint testing with reversal of hypotheses that have been presented in the literature: classical ideas, joint confirmation (Charemza and Syczewska, 1998, CS), and Bayes tests. We review the problem as it is viewed in the more classical (see, for example, Lehmann and Romano, 2005) or in the more Bayesian tradition (see, for example, Ferguson, 1967, or Pratt *et al.*, 1995).

Section 3 highlights the differences across the three approaches graphically and generally. In Section 4, we consider an application in time-series analysis, viz. the statistical testing problem that was considered by CS and Keblowski and Welfe (2004). A decision is searched for the existence of a unit root in a time-series process on the basis of two tests. One of them has the existence of a unit root as its null hypothesis and its non-existence as its alternative, while the roles of null and alternative are reversed for the second hypothesis test. Section 5 concludes.

2 Pairs of tests with role reversal of hypotheses

2.1 The general decision problem

We consider statistical decision problems of the following type. The maintained model is expressed by a parameterized collection of densities. The parameter space Θ is possibly infinite-dimensional. Based on a sample of observations from an unknown member $\theta \in \Theta$, a decision is searched on whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, with $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The event $\{\theta \in \Theta_0\}$ is called the *null hypothesis*, while $\{\theta \in \Theta_1\}$ is called the *alternative hypothesis*. If $\theta \in \Theta_0$, 'the null hypothesis is correct', while for $\theta \in \Theta_1$ the 'alternative is correct'. While the parameter space Θ may be infinite-dimensional, classification to Θ_0 and Θ_1 should rely on a finite-dimensional subspace or 'projection'. If the finite-dimensional projection of θ is observed, θ can be allotted to Θ_0 or Θ_1 with certainty. The occurrence of non-parametric nuisance is crucial for the problem that we have in mind.

Example. An observed variable X is a realization of an unknown real-valued probability law with finite expectation. Θ_0 may consist of those probability laws that have $\mathrm{E}\, X = 0$, while Θ_1 may be defined by $\mathrm{E}\, X \neq 0$. Decision is searched for a one-dimensional parameter, while Θ is infinite-dimensional.□

A characteristic feature of statistical decision problems is that θ is not observed. If θ were observed, this would enable perfect classification. This perfect case can be envisaged as incurring zero loss, in line with the usual concept of the *loss* incurred by a decision. Typically, a sample of observations for a random variable X is available to the statistician, where the probability law of the random

variable is governed by a density $f_\theta(.)$. In many relevant problems, observing an infinite sequence of such observations allows to determine θ almost surely and therefore to attain the loss of zero that accrues from direct observation of θ.

Typically, finite samples will imply non-zero loss. Following classical tradition, incorrect classification to Θ_0 is called a *type II error*, while incorrect classification to Θ_1 is called a *type I error*. For many decision problems, testing procedures can be designed that take both *type I* and the *type II* errors to zero probability, as the sample grows to infinity. We note, however, that hypothesis tests with 'fixed significance level' do not serve this aim.

2.2 Decisions in the classical framework

Assume τ_1 is a test statistic for the decision problem with the null hypothesis $\theta \in \Theta_0$ and the alternative $\theta \in \Theta_1$, while τ_2 is a test statistic with reversed null and alternative hypotheses. A hypothesis test using τ_1 will usually be designed to have a pre-assigned upper bound α for the probability of a type I error $\mathcal{P}_1(\theta)$, such that $\mathcal{P}_1(\theta) \leq \alpha$ for all $\theta \in \Theta_0$. Furthermore, the test will be designed such that the probability of a type II error $\mathcal{P}_2(\theta)$ will be minimized in some sense for $\theta \in \Theta_1$. While $\mathcal{P}_2(\theta)$ will critically depend on $\theta \in \Theta_1$ in finite samples, test consistency requires $\mathcal{P}_2(\theta) \to 0$ as $n \to \infty$ for every $\theta \in \Theta_1$.

By construction, the error probabilities for the test defined by τ_2 will have reversed properties. Therefore, a decision based on the two individual tests and common α will have a probability of incorrectly selecting Θ_0 bounded by α, and the same will be true for the probability of incorrectly selecting Θ_1. First assume independence of the two test statistics. Then, for some parameter values, these error probabilities may be close to $\alpha(1-\alpha)$, while for others they may be much lower. If both individual tests are consistent, both error probabilities should converge to zero, as the sample size increases. Thus, the joint test achieves full consistency in the sense of both $\mathcal{P}_1(\theta) \to 0$ for $\theta \in \Theta_0$ and $\mathcal{P}_2(\theta) \to 0$ for $\theta \in \Theta_1$. Even allowing for some dependence of the two test statistics are dependent will not invalidate the argument. Full consistency, which is not typical for classical tests, comes at the price that the true significance level of the test is less than α for all sample sizes.

A drawback is that the decision for Θ_0 is implied only if the test based on τ_1 'fails to reject' and the test based on τ_2 'rejects'. If the τ_1 test 'rejects' and the τ_2 test 'fails to reject', a decision for Θ_1 is suggested. In cases of double rejection or double non-rejection, no coercive decision is implied. Allotting these parts of the sample space arbitrarily to the Θ_0 or the Θ_1 decision areas would express a subjective preference toward viewing the hypothesis design of one of the two individual tests as the correct one and, therefore, the other test design as 'incorrect'.

Some, with HATANAKA(1996), interpret conflicting outcomes as indicating invalidity of the maintained hypothesis Θ. While this may be plausible in some

problems, it may require an approximate idea of possible extensions of the maintained model $\Theta^e \supset \Theta$. Clearly, in a stochastic environment, contradictory outcomes will not have a probability of zero, whatever has been the data-generating model. We assume that a complete decomposition of the sample space into two regions Ξ_0 (preference for Θ_0) and Ξ_1 (preference for Θ_1) is required. This can be achieved by basing the choice on comparing the p–values of individual tests. If both individual tests reject, the rejection with the larger p–value is ignored. Similarly, if both tests 'do not reject', the lower p–value is taken as indicating rejection. It appears that this casual interpretation of p–values is quite common in practice. By construction, the decision rule ignores any dependence among τ_1 and τ_2.

2.3 Joint confirmation

Let the acceptance regions of the two tests using τ_1 and τ_2 be denoted as Ξ_0^1 and Ξ_1^2, and similarly their rejection regions as Ξ_1^1 and Ξ_0^2. Then, one may consider basing the decomposition (Ξ_0, Ξ_1) on bivariate intervals. It is straight forward to allot the 'clear' cases according to

$$\begin{aligned} \Xi_0^1 \cap \Xi_0^2 &\subset \Xi_0, \\ \Xi_1^1 \cap \Xi_1^2 &\subset \Xi_1. \end{aligned} \tag{1}$$

In the remaining parts of the sample space, the two statistics seemingly point to different conclusions. Seen as tests for 'null' hypotheses Θ_0 and Θ_1, allotting these parts to Ξ_0 or Ξ_1 may result in 'low power' or in violating the 'risk level' condition.

'Joint confirmation hypothesis' testing or 'confirmatory analysis', according to CS, targets a *probability of joint confirmation* (PJC), which is defined as 'deciding for Θ_1, given the validity of Θ_1'. Consider the error integrals

$$\mathcal{P}_1(\theta) = \int_{\Xi_1} f_\theta(x)\, dx, \quad \mathcal{P}_2(\theta) = \int_{\Xi_0} f_\theta(x)\, dx, \tag{2}$$

Let us view the joint test as having Θ_0 as its 'null' and Θ_1 as its 'alternative'. For $\theta \in \Theta_1$, $\mathcal{P}_2(\theta)$ is the probability of a type II error, while, for $\theta \in \Theta_0$, $\mathcal{P}_1(\theta)$ is the probability of a type I error. CS define the PJC as $\mathcal{P}_1(\theta)$ for some $\theta \in \Theta_1$. Since $\mathcal{P}_2(\theta) = 1 - \mathcal{P}_1(\theta)$, the PJC simply is one minus the type II error probability for a specific $\theta \in \Theta_1$. The error integral $\mathcal{P}_2(\theta)$ is evaluated for some θ, which are members of the *alternative* for the test construction τ_1 and members of the *null hypothesis* for the construction of τ_2. Therefore, $\mathcal{P}_2(\theta)$ expresses the probability that τ_1 would wrongly accept its null and τ_2 would correctly accept its null if the tests were used individually. If (τ_1, τ_2) is used jointly, it is the probability of an incorrect decision for some given $\theta \in \Theta_1$.

Usually, there is a manifold of pairs (τ_{a1}, τ_{a2}) such that $(\tau_1, \tau_2) = (\tau_{a1}, \tau_{a2})$ implies the condition $\mathcal{P}_1(\theta) = 1-\alpha$ for a given α. Among them, joint confirmation selects critical points (τ_{c1}, τ_{c2}) by the condition that $\mathcal{P}_1(\theta)$ coincide for the two component tests that build on individual τ_j and corresponding τ_{cj}. While this superficially looks like a Bayesian critical point, where the probabilities of Θ_0 and Θ_1 coincide, no probability of hypotheses is used, as the procedure is built in the classical way, where hypotheses do not have probabilities. While an informal Bayesian interpretation of p –values may interpret them as such probabilities, a genuine Bayes test determines critical points by comparing probabilities for Θ_0 and Θ_1, not two measures of probability for Θ_1. An apparent advantage of joint confirmation is, however, that it avoids the Bayesian construction of weighting functions.

2.4 Bayes tests

The occurrence of an apparent contradiction by two individual hypothesis tests has a relatively simple solution in Bayesian statistics. The admissible parameter space is defined as $\Theta_0 \cup \Theta_1$ and the remainder is *a priori* excluded, according to the statement of the decision problem. After fixing weight functions h_0 and h_1 on the hypotheses and a loss criterion g, the decision problem can be subjected to computer power and yields a solution that is optimal within the pre-defined set of admissible decompositions (Ξ_0, Ξ_1). For example, one may restrict attention to decompositions that are based on the test statistics (τ_1, τ_2) and on bivariate intervals.

The Bayesian setup to testing problems assumes weighting functions h_0 and h_1 on the respective parameter spaces Θ_0 and Θ_1, which can be interpreted as probability densities. While a usual interpretation of h_0 and h_1 is that they represent *a priori* probabilities of parameter values, it is not necessary to adopt this interpretation for Bayes testing. If the sample space, for example $\mathbb{R}^n$ for sample size n, is partitioned into two mutually exclusive subsets Ξ_0 and Ξ_1, such that $X \in \Xi_j$ implies deciding for $\theta \in \Theta_j$, the probability of a *type I* error is $\mathcal{P}_1(\theta)$ for a given member $\theta \in \Theta_0$. The Bayes weighting scheme allows to evaluate

$$\mathcal{L}_1(h_0, \Xi_1) = \int_{\Theta_0} \int_{\Xi_1} f_\theta(x)\, dx\, h_0(\theta)\, d\theta \tag{3}$$

as a measure for the 'average' *type I* error involved in the decision. Conversely, the integral

$$\mathcal{L}_2(h_1, \Xi_0) = \int_{\Theta_1} \int_{\Xi_0} f_\theta(x)\, dx\, h_1(\theta)\, d\theta = \int_{\Theta_1} \mathcal{P}_2(\theta)\, h_1(\theta)\, d\theta \tag{4}$$

represents the 'average' *type II* error involved. A Bayesian view of the decision problem is to minimize the *Bayes risk*

$$g\left(\mathcal{L}_1(h_0, \Xi_1), \mathcal{L}_2(h_1, \Xi_0)\right)$$

$$= \quad g\left(\int_{\Theta_0}\int_{\Xi_1} f_\theta(x)\,dx\,h_0(\theta)\,d\theta, \int_{\Theta_1}\int_{\Xi_0} f_\theta(x)\,dx\,h_1(\theta)\,d\theta\right) \tag{5}$$

in the space of possible partitions of the sample space, for a given function $g : \mathbb{R}^2 \to \mathbb{R}^+$. The function g is designed to express the afore-mentioned loss. Therefore, $g(0,0) = 0$ and monotonicity in both arguments are useful restrictions. If for any $\theta \in \Theta_j$ no observed sample generated from that θ implies the incorrect decision Ξ_k with $k \neq j$, both arguments are zero and the loss is zero. Zero Bayes risk can also be attained if incorrect decisions occur for subsets $\tilde{\Theta}_j \subset \Theta_j$ with $h_j = 0$ or $\int_{\tilde{\Theta}_j} h_j d\theta_j = 0$ only.

By construction, Bayes tests attain full consistency $\mathcal{P}_2(\theta) \to 0$ for $\theta \in \Theta_1$ and $\mathcal{P}_1(\theta) \to 0$ for $\theta \in \Theta_0$ by minimizing $g(\mathcal{L}_2(h_0, \Xi_1), \mathcal{L}_1(h_1, \Xi_0)) \to 0$, except on null sets of the measure that is defined by the weighting priors h_0, h_1. In a Bayesian interpretation, the classical approach is often viewed as allotting a strong relative implicit weight to Θ_1 in smaller samples, which makes way to a strong weight on Θ_0 as the sample size grows. This 'weight' refers to the corresponding derivatives g_1 and g_2 of the g function, not to the weighting priors h_0 and h_1.

3 A graphical comparison of three methods

Assume the individual test using τ_1 rejects in the left tail of the range, while the test using τ_2 rejects in the right tail.

For an instructive comparison across the methods, first consider Figure 1. Without restricting generality, we consider a situation where Θ_0 corresponds to the null of the test using τ_1 and to the alternative of the test using τ_2. It is convenient to 'code' both tests in terms of their respective null distributions. In other words, the diagram is not drawn in the original coordinates $(\tau_1, \tau_2) \in \mathbb{R}^2$ but rather in the fractiles $(F_1(\tau_1), F_2(\tau_2)) \in [0,1]^2$ for distribution functions F_1 and F_2. Because distribution functions are monotonous transforms, the information remains identical for any such functions. However, it eases the interpretation of the diagrams if F_j corresponds to the 'null distributions' of τ_j, assuming that a distribution of τ_j under its null hypothesis is (approximately) unique. In short, we label the axes by $F_1(\tau_1)$ and $F_2(\tau_2)$. Rejection by the test using τ_1 corresponds to $F_1(\tau_1) < \alpha$. In the following, we adopt the conventional level $\alpha = 0.05$. If the test using τ_1 rejects and the test using τ_2 does not or the sample is in $\Xi_1^1 \cap \Xi_1^2$, there is clear evidence in favor of hypothesis Θ_1. Conversely, if the first test does not reject and the second test does so or the sample is in $\Xi_0^1 \cap \Xi_0^2$, we have clear evidence in favor of Θ_0. If both tests accept or reject, the evidence remains unclear. This fact is expressed by leaving the north-west and south-east regions white.

The diagonal line indicates the informal classical solution of allotting the undecided regions according to a simple comparison of p–values. If this procedure

is adopted, all points above the line are interpreted as preferring Θ_0 and those below the line as preferring Θ_1.

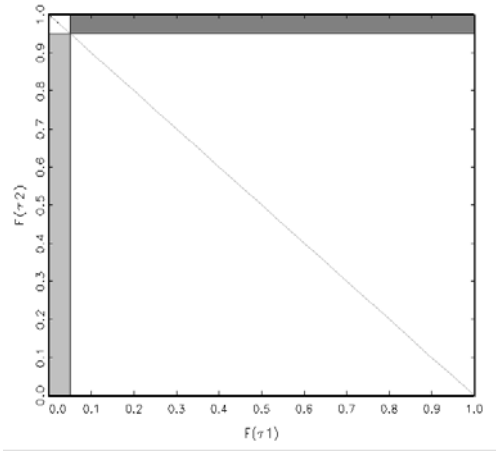

Figure 1: Classical decision following the joint application of two classical tests with switching null hypotheses. Axes are determined by the null distribution of τ_1 and the null distribution of τ_2. Light gray area represents decisions in favor of Θ_1, while the dark gray area corresponds to Θ_0.

Next, consider Figure 2. It represents the decision of joint confirmation. Rather than using the null distributions of the two test statistics τ_1 and τ_2, we use here the null distribution of τ_1 but the alternative distribution of τ_2 and code the two test statistics accordingly. Usually, *the* alternative distribution does not exist, therefore one uses a representative element from the τ_2 alternative. If τ_2 rejects *and* τ_1 accepts, this is the 'confirmation area' of hypothesis Θ_0. Its probability under the representative distribution from Θ_0 has a given probability α. Along the $(x, 1-x)$–diagonal, individual rejection probabilities coincide, thus the corner point is selected.

A possible interpretation of the method's focus on the north-east confirmation area is that the dark gray area favors Θ_0, while the remaining area favors Θ_1. The work of CS appears to support this interpretation by using a similar coloring of the four areas in a histogram plot. The interpretation is not coercive, however, and one may also forego a decision in conflicting cases, as in the classical rule of Figure 1. Then, joint confirmation becomes closer in spirit to reversing a classical test by replacing the original alternative by a point alternative, such that it becomes a convenient null. We refrain from this simplifying view, which is invalid in the classical tradition, and we view the joint confirmation decision according to Figure 2. In any case, the procedure is asymmetric, as confirming Θ_1 leads to a different solution from confirming Θ_0. The choice of confirmed

hypothesis is not entirely clear. CS and KEBLOWSKI AND WELFE (2004) choose the null of the more popular component test.

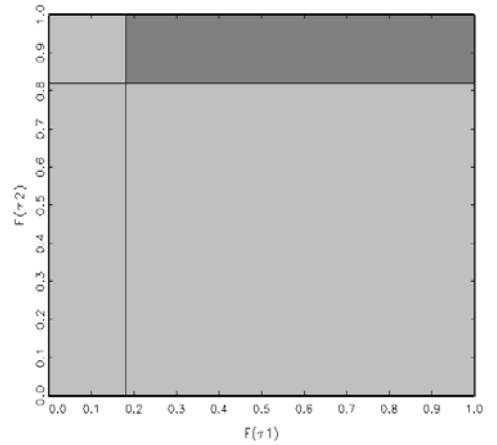

Figure 2: Joint-confirmation decision following the joint application of two classical tests with switching null hypotheses. Axes are determined by the null distribution of τ_1 and a representative alternative distribution of τ_2. Light gray area represents decisions in favor of Θ_1, while the dark gray area corresponds to Θ_0.

A typical outcome of a Bayes test is depicted in Figure 3. As in the classical test in Figure 1, axes correspond to respective null distributions functions $F_j(x) = \int_{-\infty}^{x} f_j(z)\,dz$. However, instead of a fixed density $f_j(z)$, we now use a weighted average $\int_{\Theta_j} f_\theta(z)\,h_j(\theta)\,d\theta$ of *all possible* null densities. Then, a simulation with 50% Θ_0 and 50% Θ_1 distributions is conducted, where all kinds of representatives are drawn, according to weight functions h_0 and h_1. Accordingly, a boundary can be drawn, where both hypotheses occur with the same frequency. Northeast of this *decision contour*, the hypothesis Θ_0 is preferred, while to the southwest the hypothesis Θ_1 is preferred. While the decision rests on a more informative basis than in the other approaches, the position of the curve is sensitive to the choice of h_0 and h_1. In a fully Bayesian interpretation, the decision contour is defined as the set of all points $\tau_c = (\tau_{c1}, \tau_{c2}) \in \mathbb{R}^2$ where $P(\Theta_0|\tau_c) = P(\Theta_1|\tau_c)$, if Θ_j have prior distributions of equal probability across hypotheses, i.e. $P(\Theta_0) = P(\Theta_1)$, and the elements of the two hypotheses have prior probabilities according to the weight functions h_0 and h_1. In the interpretation of the decision framework that we introduced in Section 2, the decision contour is the separating boundary of the two regions Ξ_0 and Ξ_1, conditional on the restrictions that only such separations of the sample space are permitted that depend on the observed statistic τ_c *and* on a loss function $g(.,.)$ that gives equal weight to its two arguments, such as $g(x,y) = x + y$.

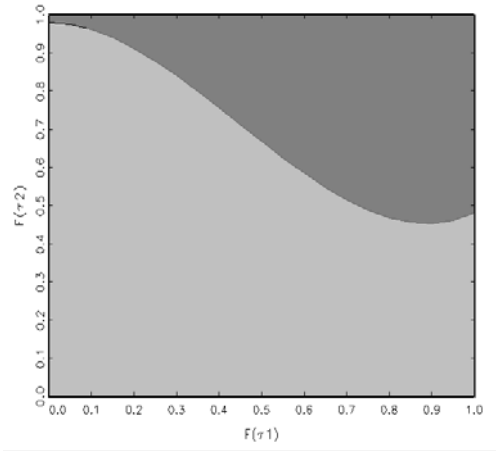

Figure 3: Bayes-test decision following the joint application of two classical tests with switching null hypotheses. Axes are determined by weighted averages of null distributions of τ_1 and τ_2. Light gray area represents decisions in favor of Θ_1, while the dark gray area corresponds to Θ_0.

The choice of h_0 and h_1 is undoubtedly important for the Bayes test, as are all types of prior distributions for Bayesian inference. There are several prescriptions for 'eliciting' priors in the Bayesian literature. To some researchers, elicitation should reflect true prior beliefs, which however may differ subjectively and are maybe not good candidates for situations with strong uncertainty regarding the outcome. Other researchers suggest to standardize prior distributions and, consequently, weight functions according to some simple scheme. Particularly for Bayes testing aiming at deriving decision contours, it appears to be a good idea to keep the weight functions flat close to the rival hypothesis. The tail behavior of the weight functions has less impact on the contours.

An important requirement is that the weighting priors are *exhaustive* in the sense that, for every $\theta \in \Theta_j$, any environment containing θ, $E(\theta)$, should have non-zero weight $h_j(E(\theta)) > 0$. This ensures that open environments within Θ_j appear with positive weight in $\mathcal{L}_1(h_0, \Xi_1)$ or $\mathcal{L}_2(h_1, \Xi_0)$ and, consequently, that the Bayes risk g converging to zero as $n \to \infty$ implies full consistency. Informally, exhaustiveness means that, for any distribution within Θ there is a distribution 'close' to it that can be among the simulated draws.

Another important choice is the loss function g. The function $g(x, y) = x + y$ corresponds to the Bayesian concept of allotting identical prior weights to the two hypotheses under consideration. In line with the scientific concept of unbiased opinion before conducting an experiment and in a search for 'objectivity', it appears difficult to accept loss functions such as $g(x, y) = (1 - \kappa)x + \kappa y$ with $\kappa \neq 1/2$. These functions are sometimes used in the Bayesian literature (for

example, see PRATT *et al.*, 1995) and may represent prior preferences for one of the two hypotheses. Classical tests with fixed significance levels can usually be interpreted as Bayes tests with severe restrictions on the allowed decompositions (Ξ_0, Ξ_1) and with unequal prior weights. Seen from a Bayes-test viewpoint, it appears difficult to justify this traditional approach.

4 Testing for unit roots in time series

4.1 The I(0)/I(1) decision problem

An important decision problem of time series analysis is to determine whether a given series stems from a stationary or a difference-stationary process. Stationary (or I(0)) processes are characterized by the feature that the first two moments are constant in time, while difference-stationary (or I(1)) processes are non-stationary but become stationary after first differencing. These two classes, I(0) and I(1), are natural hypotheses for a decision problem. Various authors have provided different exact definitions of these properties, thereby usually restricting the space of considered processes. For example, instead of stationary processes one may focus attention on stationary ARMA processes, and instead of difference-stationary processes one may consider accumulated stationary ARMA processes. Usually, the class I(0) excludes cases with a spectral density that disappears at zero.

This is, roughly, the framework of DICKEY AND FULLER (1979, DF) who introduced the still most popular testing procedure. Their null hypothesis Θ_0 contains finite-order autoregressive processes $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \varepsilon_t$, formally written $\phi(B)\, x_t = \varepsilon_t$ with white-noise errors ε_t and the property that $\phi(1) = 0$, while $\phi(z) \neq 0$ for all $|z| \leq 1$, excepting the one unit root. We use the notation B for the lag operator $BX_t = X_{t-1}$ and $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ for general argument z. The corresponding alternative Θ_1 contains autoregressions with $\phi(z) \neq 0$ for all $|z| \leq 1$. This is a semiparametric problem, as distributional properties of ε_t are not assumed, excepting the defining properties for the first two moments. In order to use asymptotic theorems, however, it was found convenient to impose some restrictions on higher moments, typically of order three or four. We note that the interesting part of both hypotheses is fully parametric, and that both Θ_0 and Θ_1 can be viewed as equivalent to subspaces of $\mathbb{R}^{\mathbb{N}}$. In particular, this 'interesting part' of $\Theta_0 \cup \Theta_1$ can be viewed as containing sequences of coefficients $(\phi_j)_{j=0}^{\infty}$ with the property that $\phi_0 = 1$ and $\phi_j = 0$ for $j > J$, for some J. Choosing Θ_0 as the null hypothesis is the natural choice, as it is defined from the restriction $1 - \sum_{j=1}^{\infty} \phi_j = 1 - \sum_{j=1}^{J} \phi_j = 0$ on the general space. Stated otherwise, Θ_0 has a 'lower dimensionality' than Θ_1, even though both spaces have infinite dimension by construction.

In the form that is currently widely used, the test statistic is calculated as

the t–statistic of a, $\hat{a}/\hat{\sigma}_a$, in the auxiliary regression

$$\Delta y_t = c + a y_{t-1} + \sum_{j=1}^{p-1} \xi_j \Delta y_{t-j} + u_t, \tag{6}$$

where Δ denotes the first-difference operator $1 - B$, $(a, \xi_1, \ldots, \xi_{p-1})'$ is a one-one transform of the coefficient sequence $(\phi_1, \ldots, \phi_p)'$, $a = 0$ iff $\phi(1) = 0$, p is either determined as a function of the sample size or by some empirical criterion aiming at preserving white noise u_t, and u_t is the regression error. $\hat{\sigma}_a$ is the usual least-squares estimate of the coefficient standard error. In line with the literature, we will refer to the case $p = 1$ as 'DF statistic' and to the case $p > 1$ as the 'augmented DF statistic'.

The distribution of the test statistic in Θ_0 was tabulated for finite samples under special assumptions on the ε_t distribution, while the asymptotic distribution was later expressed by integrals over Gaussian continuous random processes (see, for example, Dhrymes, 1988). In finite samples, $\mathcal{P}_1(\theta) = \alpha$ does not hold exactly for all $\theta \in \Theta_0$, and $\mathcal{P}_0(\theta) \leq 1 - \alpha$ will not hold for all $\theta \in \Theta_1$. Straightforward application of the test procedure with fixed α will result in $\mathcal{P}_1(\theta) \to \alpha$ for $\theta \in \Theta_0$ and $\mathcal{P}_0(\theta) \to 0$ for $\theta \in \Theta_1$ if $n \to \infty$.

While the decision model was introduced in the purely autoregressive framework by DF—such that eventually $p \geq J$ and $u_t = \varepsilon_t$—, it was extended to ARMA models by later authors. In other words, the test statistic continues to be useful if u_t is MA rather than white noise, assuming some further restrictions, such as excluding unit roots in the MA polynomial. Several authors studied the properties of the DF test *outside* $\Theta_0 \cup \Theta_1$. For example, it was found of interest to investigate the cases that the polynomial zero under I(1) has a multiplicity greater than one (see Pantula, 1989) and that the processes have some simple features of non-stationarity under both I(0) and I(1) (see Perron, 1989, and Maddala and Kim, 1998).

A different test for the unit roots problem is the KPSS test (after Kwiatkowski *et al.*, 1992). According to Tanaka (1996), its test statistic has the appealingly simple form

$$K = n^{-1} \frac{y' M C C' M y}{y' M y}, \tag{7}$$

where y is the vector of data, M corrects for means or trends, and C is used to accumulate a series in the sense of the operator Δ^{-1}. This version is correct for testing a null hypothesis that y is white noise against some alternative where y is a random walk, which is a most unlikely situation. If the null hypothesis is to contain more general stationary processes, KPSS suggest a non-parametric correction of the above statistic. The correction factor r is defined as $r = \tilde{\sigma}_S^2 / \tilde{\sigma}_L^2$, where $\tilde{\sigma}_S^2$ is an estimate of the variance of a mean-corrected version of Δy, $\eta = \Delta y - m_{\Delta y}$ for $m_{\Delta y} = (n-1)^{-1} \sum_{t=2}^{n} \Delta y_t$, and $\tilde{\sigma}_L^2$ is an estimate of the re-scaled

zero-frequency spectrum of the same process. We follow the specification for these estimates as

$$\begin{aligned}\tilde{\sigma}_S^2 &= n^{-1}\sum_{t=2}^{n}\eta_t^2, \\ \tilde{\sigma}_L^2 &= \tilde{\sigma}_S^2 + 2n^{-1}\sum_{j=1}^{l}\left(1-\frac{j}{l+1}\right)\sum_{t=j+2}^{n}\eta_t\eta_{t-j}. \end{aligned} \tag{8}$$

This is a Bartlett-type estimate of the frequency-zero spectrum. We follow one of the suggestions in the literature for specifying the upper bound l as the integer part of $5\sqrt{n}/7$. For a comparison, we consider the uncorrected version K as well as the corrected version $\tilde{K} = rK$. We start by presenting our results for K.

It is to be noted that the null hypothesis of the DF test and the alternative of the KPSS test do not match exactly. The same is true for the DF alternative and the KPSS null and for the maintained models of both tests. Numerous studies aimed at generalizing the hypotheses particularly for the DF test, thus it is not easy to determine the universally accepted maintained model and null hypotheses for both cases. It appears that the KPSS model can be transformed into an ARMA representation with a unit root in the autoregressive polynomial, which cancels under the KPSS null with a unit root in the moving-average polynomial. Tests for moving-average unit roots have been developed by Saikkonen and Luukkonen (1993), among others, and they have comparable null distributions. By contrast, the maintained model of the DF test is ARMA, with the case of moving-average unit roots excluded. Due to different parameterizations of the infinite-dimensional maintained hypothesis space, it is conceivable that the same hypothesis can serve as a null in one test and as an alternative in another test, with restrictions on single parameters defining null hypotheses in both cases. Stock (1994) presented a fully consistent classification that relies on a single test statistic, which is a variant of $\tilde{K}$, and mentioned that a similar aim might be achieved by using the DF statistic. Recently, Mueller (2004) showed that the construction of a consistent discrimination test with a very general version of an $I(1)$ alternative is not possible. For practical purposes, one may restrict attention to hypotheses Θ_0 and Θ_1 that make the two tests consistent on their alternatives and that are simultaneously reasonably representative of 'I(0)' and 'I(1)'.

Keblowski and Welfe (2004) give a large-n 5% point, according to their construction, at -3.10 for the DF test and at 0.42 for the KPSS test. Both tests have regions pointing to stationarity in the left tails of their null distribution. Their classical interpretation differs across the two tests. For the DF test, I(1) is the null hypothesis, and rejection in the left tails indicates stationarity. For the KPSS test, stationarity is the null hypothesis, and rejection in the right tails indicates I(1). Therefore, in classical testing, secure decisions are only taken in

the areas $[0, \alpha] \times [0, 1-\alpha]$ and $[\alpha, 1] \times [1-\alpha, 1]$ of the coded null fractiles table that we used in Figure 1. Using the values -3.10 and 0.42, KEBLOWSKI AND WELFE determine an acceptance region for I(1) as $(-3.10, \infty) \times (0.42, \infty)$, thus suggesting to decide for the stationarity hypothesis in the remainder of $\mathbb{R}^2$. This situation could be drawn in a re-coded version just as in Figure 2.

4.2 Bayes-test experiments

In order to provide the Bayesian test solution, we consider parameterizing the two hypotheses Θ_0 and Θ_1. The stationary processes are first-order autoregressions $X_t = \phi X_{t-1} + \varepsilon_t$ ('AR(1)'), with ϕ distributed uniformly on $(-1, 1)$. The integrated processes are random walks $X_t = X_{t-1} + \varepsilon_t$. The errors ε_t are drawn independently from a standard normal distribution. All trajectories are started from zero and have length $n = 100$. This design may bias the results in favor of the DF test, as it corresponds to its construction.

After setting up the Bayesian weighting framework, defining implicitly h_0 and h_1, trajectories are drawn and statistics, (non-augmented) DF and (uncorrected) KPSS, are being calculated for each trajectory. We chose the replication size of 2×10^6, meaning that 10^6 random-walk and 10^6 stationary trajectories were used. The DF statistics from the random walks and the KPSS statistics from the stationary autoregressions define the null distributions for further steps. While the DF distribution corresponds well to its tabulated and also to its asymptotic form, the KPSS null distribution cannot correspond to its theoretical one, which rests on white-noise trajectories. It is informative to draw the empirical distribution functions for null and alternative models. In the case of the DF statistic, the distribution functions have similar shapes and show a satisfactory discrepancy among them. This means that the average alternative model for the DF test may really correspond to white noise, as it should according to construction, while the drawn trajectories of course contain negatively correlated specimens as well as near random walks. We obtain a different picture for the KPSS statistic, where the alternative distribution has an almost uniform appearance. However, also for the KPSS statistic, there is a comforting difference in shape between the average null and the average alternative distribution.

The next step is setting up a grid in the fractile space. We chose a 100×100 grid, which corresponds to an average of 200 entries per bin. It turned out that many bins are indeed empty and that the smoothness of the boundary curve is not quite satisfactory. In this situation, increasing replications or kernel smoothing are possible options. One may also consider reducing the resolution of the grid. While the coordinates for the plot are dictated by the two null distribution functions, we provide the more informative values of the fractiles on both axes, instead of labeling them simply by the values of the distribution functions, as in Figures 1–3.

Then, the grid is filled with the simulated statistics, the origin of which is

known, according to whether they belong to Θ_0 or to Θ_1. Bins with a preponderance of stationary processes are interpreted as suggesting a decision in favor of Θ_0, while other bins are allotted to Θ_1. The separation of the sample space, as coded by null distributions of the two test statistics, into Ξ_0 and Ξ_1, as it were, is shown in Figure 4. The boundary or *decision contour* runs south-east from the north-west corner. The large light gray area in the south-east marks test outcomes that were not observed in the simulation. The probability—in the Bayesian sense—of large or even average DF statistics *together* with small or even average KPSS statistics is very low. Supports remain unbounded, and an unreasonably large number of replications will succeed in determining the decision contour in the south-east fields. It is questionable whether the exercise is worth the computer time, as these values were not generated with a reason. They are simply very unusual in empirical practice. In order to corroborate this statement, one might require extending Θ_0 and Θ_1 to cover more general stationary and integrated processes. We will point at the consequences in some more sophisticated experiments below.

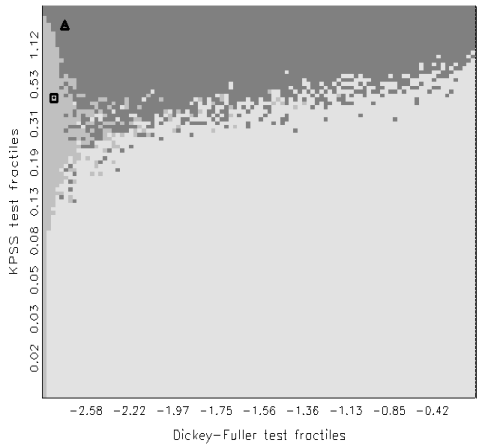

Figure 4: Bayes-test decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. Sample size is $n = 100$. Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

Note that the boundary is much more informative than any of the alternative decision concepts in classical statistics. For example, one sees that the DF statistics show a much sharper concentration under their stationary alternative than the KPSS statistics under their random-walk alternative. This feature

may reflect the missing correction for serial correlation in the uncorrected KPSS statistic K. One also obtains particular information on the behavior in the northwest corner of conflict, where DF statistics point to stationary behavior, while KPSS statistics indicate unit roots. The decision contour deviates from a straight $(0,1)-(1,0)$ diagonal and appears to emphasize the role of the KPSS statistic in conflict situations. In summary, the graph allows a truly bivariate evaluation of the information provided by the two test statistic, which the classical approaches do not. For a comparison, Figure 4 displays the vertices for a classical 5% test by a triangle and for joint confirmation by a square.

Table 1: Probability of incorrect decisions.

design	Bayes	joint conf.	p-value rule	classical r.
autoregressive	0.053	0.057	0.060	0.094
moving average	0.005	0.016	0.010	0.028
autoregressive, $n = 20$	0.173	0.193	0.182	0.273
AR, modified KPSS	0.056	0.058	0.092	0.157
state space	0.024	0.062	0.043	0.112
AR + state-space	0.101	0.235	0.130	0.295
AR + state-space, augm. DF	0.077	0.095	0.088	0.144

The numbers in Table 1 allow a cursory assessment of the different approaches. Bayes risks were evaluated for the Bayes test, for the joint-confirmation technique, and for two classical variants. We assume a loss of one for an incorrect classification and a loss of zero for a correct one, i.e. a zero-one loss. For the Bayes test, classification rests on the preference area, exactly as shown in the graphs. This is not plausible and biases the results in favor of the Bayes test, as a typical user would try to smooth the decision contour. Therefore, the value for the Bayes test serves as a lower bound or as a benchmark. For joint confirmation, we rely on the critical values tabulated by Keblowski and Welfe (2004) and decide according to a diagram as in Figure 2. For the p–value approach, we assume a decision relying on a negative diagonal as in Figure 1. Finally, for a worst-case benchmark, we interpret the classical approach as taking a decision only in the areas with a unanimous suggestion and as randomizing the decision otherwise. Clearly, the last solution is worse than any competitor, with an error probability close to 10%. Joint confirmation beats p–value checks, while it comes close to the Bayes-test benchmark.

To assess the sensitivity of the boundary curve, another experiment generates Θ_0 from moving-average processes with uniform $\theta \in (-1,1)$ and $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$. Then, Θ_1 is not contained in the topological closure of Θ_0, and discriminating the two hypotheses becomes easier. In Figure 5, we see that the boundary moves west, due to an extreme concentration of the distribution of simulated DF test

statistics along the western border. The graph suggests using the DF test at a significance level of 1%. In case of rejection, the observed time series should be stationary. In case of no rejection, the KPSS statistic will be 'large' and the time series will stem from an integrated process. We note that these recommendations only make sense if we know that any potential stationary process is first-order moving-average, which is a most unlikely situation.

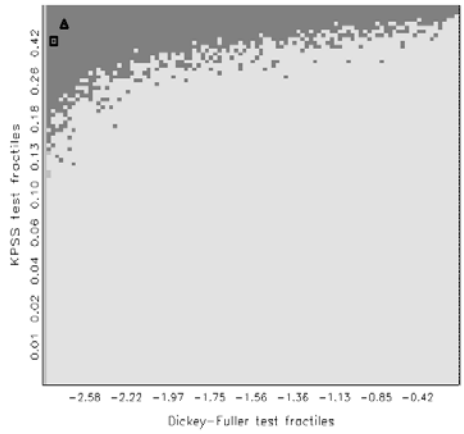

Figure 5: Bayes-test decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. Sample size is $n = 100$. Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order moving-average processes with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

Table 1 shows that error probabilities have indeed decreased impressively. We note that the p–value decision dominates joint confirmation.

In order for the Bayes method to be useful to empirical researchers, some standardization of flexible features like h_0 and h_1 will be necessary. While such a standardization of weighting priors will hardly satisfy the convinced Bayesian, it appears possible and should correspond to the ubiquitous 5% of classical statistics. For example, in this example the design of h_0 in Figure 5 is unsatisfactory, while the one in Figure 4 is 'better'.

Let us re-consider the critical point $(-3.10, 0.42)$ that was provided by Keblowski and Welfe (2004) as a solution to the joint confirmation approach. In the original coordinates of Figure 4, the vertex is shown as a square approximately at $(0.03, 0.76)$. The region $(-3.10, \infty) \times (0.42, \infty)$ in coordinates of (τ_1, τ_2) and $(0.03, 1) \times (0.76, 1)$ in the diagram coordinates belongs to the 'confirmation

area' for I(1). Thus, a large part of the confirmation area coincides with the hypothesis-I(1) area of the Bayes test. Test outcomes in the north-west to the joint-confirmation point, however, would be classified differently. Test outcomes in the south-east are rare, and their classification is of little empirical relevance. The approximate coincidence of decisions is not a systematic property of the two procedures, and it depends on the sample size. The choice of weighting functions for the Bayes procedures is not unique, hence the coordinate system looks differently for different h_0 or h_1. For example, if draws for the I(0) hypothesis are restricted to the textbook case of white noise, instead of extending them to autoregressive or moving-average processes, the point $(-3.10, 0.42)$ will appear as $(0.03, 0.94)$, which comes closer to the construction idea of joint confirmation. All decision maps for the decision suggested by joint-confirmation analysis have the typical shape of Figure 2, with the vertex points indicated as squares in the Bayes diagrams.

As the sample size increases, the distribution of test statistics in the Bayesian procedure converges to the western—for I(0)—and northern—for I(1)—borders, and the decision contour disappears in the north-west corner. By contrast, the decisions of classical and of joint-confirmation statistics rely on fixed respective critical points close to $(0.05, 0.95)$ or $(0.03, 0.76)$. Convergence of joint-confirmation critical points to their limits is conveniently fast, as can be seen from the tables provided by CS or Keblowski and Welfe (2004).

Figure 6 allows an impression of the influence of the sample size on decision contours. It differs from the experiment of Figure 4 by using $n = 20$ instead of $n = 100$. In such small samples, the central area is reasonably populated, and the decision contour spreads along the $y = 1 - x$ diagonal, although with a slightly shifted position. Apparently, 'rejection' according to the KPSS statistic is given priority, such that the north-west corner is in the hands of the I(1) hypothesis. Table 1 shows that error probabilities of all procedures are similar, excepting the classical rule with randomized decision. Because the decision contour is close to the negative diagonal, the p–value rule comes close, as it uses that diagonal as an exact decision contour.

If the uncorrected KPSS statistic K is replaced by the statistic $\tilde{K}$, which according to theoretical results is more appropriate in our design, we obtain the contour plot in Figure 7. It is obvious that the influence of the correction term is strong. Convergence of the finite-sample distribution of $\tilde{K}$ to its limit is much slower than for the uncorrected K. The decision contour is now recognizable for a larger part of the diagram. The picture suggests to rely mainly on the decision suggested by the DF test. Time series with DF statistics that do not imply individual rejection are likely to be integrated, even when $\tilde{K}$ is moderately low. The diagram does not imply that $\tilde{K}$ is not a useful statistic, as the simulation design favors the parametric DF test. Neither does it imply that K should be used instead of $\tilde{K}$. However, it suggests that, assuming time series have actually been generated from first-order autoregressions, low $\tilde{K}$ values observed together

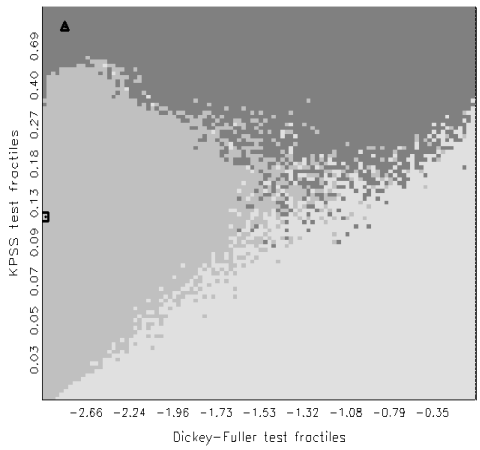

Figure 6: Bayes-test decision following the joint application of the Dickey-Fuller and KPSS tests with switching null hypotheses. Sample size is $n = 20$. Axes are determined by weighted averages of null distributions of the Dickey-Fuller and KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

with inconspicuous DF values point to non-stationary generating processes. This information may be valuable and it is in outright contradiction to the classical *and* to the joint-confirmation approaches. With the joint-confirmation vertex in the extreme left tail of the DF null distribution, the joint-confirmation decision essentially is based on a pure $\tilde{K}$ evaluation, which appears inefficient. The visual impression is confirmed by the Bayes-risk evaluation in Table 1. The difference between the Bayes risk due to Bayesian decision contours and joint confirmation is substantial, and the classical procedures are not competitive.

It was pointed out before that the sampling design gives an advantage to the parametric procedure, as the generating model corresponds to the testing model of the DF test but not to the testing model of the KPSS test. In order to remove that advantage, we now consider a mixed generating design, where 50% of the I(0) and I(1) processes are generated as before. The remaining 50% are generated from sums of stationary first-order autoregressions and random walks in the form

$$\begin{aligned} y_t &= x_t + u_t, \\ x_t &= x_{t-1} + \xi_t, \\ u_t &= \phi u_{t-1} + \varepsilon_t. \end{aligned} \tag{9}$$

The stationary x_t process is generated as AR(1) as before, with ϕ uniformly drawn from $U(-1, 1)$ and ε_t drawn from $N(0, 1)$. The main difference is the $N(0, \sigma^2)$

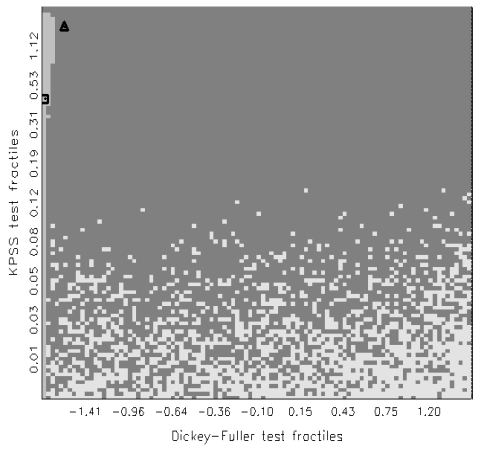

Figure 7: Bayes-test decision following the joint application of the Dickey-Fuller and corrected KPSS tests with switching null hypotheses. Sample size is $n = 100$. Axes are determined by weighted averages of null distributions of the Dickey-Fuller and corrected KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

process ξ_t. For the $I(0)$ hypothesis, $\sigma^2 = 0$, while for the $I(1)$ hypothesis σ^2 is drawn from a standard half-normal distribution. We note that the $I(0)$ design is the same as before, such that 50% of all processes are generated from that model, while 25% obey the pure random walk null of the DF model and 25% obey the mixed state-space model (9). A similar concept was adopted by Kunst and Reutter (2002) in their evaluation of statistics for decisions on seasonal behavior.

A simulation experiment with 100% trajectories taken from the state-space design (9) is reported in Table 1. Joint confirmation yields a higher Bayes risk than the p–value rule, as the average I(1) process generated from the design differs from the random walk that underlies the joint-confirmation points. The p–value rule is, in turn, inferior to the Bayes test, as the decision contour is nearly vertical and differs from the diagonal assumed by the p–value rule.

Of more interest is mixing the two simulation designs, as outlined above. For the Bayes test, we obtain the contour plot in Figure 8. Apparently, the decision for the DF test now changes to a value that is much closer to the null median than to the lower-tail fractiles that were observed in other charts. However, the null distribution of the DF test has now also been changed, due to a different generating model for I(1), and the shift of the DF contour to the right is much

less pronounced in real values than may be suggested by the chart. The KPSS test now shows its strength, particularly in the north-west corner. Large values of the KPSS statistic now imply an I(1) decision, even when the DF test tells otherwise. Again, joint confirmation incurs a higher Bayes risk than the p–value rule. The difference is mainly due to the north-west corner, where the Bayes test recommends I(1) and the p–value rule splits the area evenly, while joint confirmation sticks to the tabulated vertex (indicated by a square) and assigns this area to I(0). The general increase in Bayes risk shows that the mixed experimental design has made a decision more difficult. Yet, a Bayes risk of 10% can be regarded as 'good', as applications of a single classical test at 5% typically yield higher average risks. Even the worst-case randomized classical method provides a lower risk than applying one of the component tests alone. In this sense, the usage of the two statistics jointly is certainly advisable.

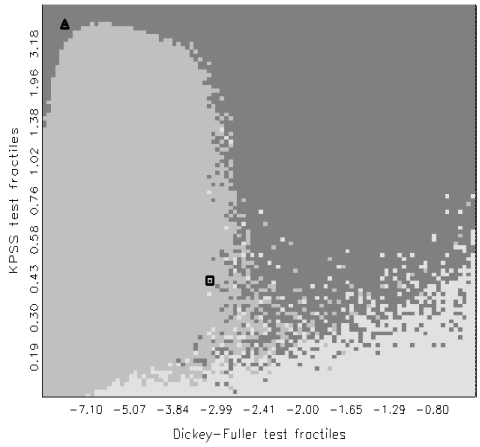

Figure 8: Bayes-test decision following the joint application of the Dickey-Fuller and corrected KPSS tests with switching null hypotheses. Sample size is $n = 100$. Sampling design corresponds to a mix of autoregressions and state-space processes. Axes are determined by weighted averages of null distributions of the Dickey-Fuller and corrected KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

The experiment of Figure 8 biases the results in favor of the KPSS test, as the KPSS statistic has been corrected for serial correlation, while the DF statistic is used naïvely. Therefore, we finally replace the DF statistic by an 'augmented' variant in the same simulation design. The augmented Dickey-Fuller test uses regression (6) with $p > 1$, thus reducing serial correlation in the errors u_t and

attaining a null distribution that comes closer to the one under the pure random-walk hypothesis. The literature recommends to determine p from the data, while we simply set $p = 2$ for the experiment. The result is shown in Figure 9. The fractiles are again close to the random-walk null and the decision contour moves back to the familiar shape. This indicates that Figure 8 is probably not representative for an advantage of KPSS testing but rather reflects an incorrect application of the DF test. We also experimented with variants of data-determined lag orders p, without any further important change in the overall shape.

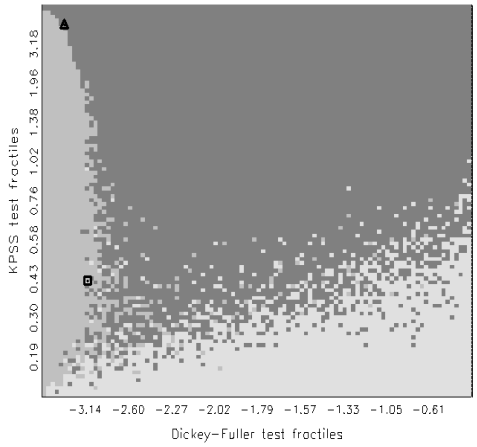

Figure 9: Bayes-test decision following the joint application of the augmented Dickey-Fuller and corrected KPSS tests with switching null hypotheses. Sample size is $n = 100$. Sampling design corresponds to a mix of autoregressions and state-space processes. Axes are determined by weighted averages of null distributions of the augmented Dickey-Fuller and corrected KPSS statistics. Stationary processes were generated from first-order autoregressions with uniform weights on the coefficients. Light gray area represents decisions in favor of stationarity, while the dark gray area corresponds to first-order integration. Very light gray corresponds to values with very low probability.

4.3 An application to economics data

For an empirical example that demonstrates the decision rules, we use consumer price inflation for three European countries: Austria, Germany, and France. Quarterly observations on the consumer price index are available in the OECD Main Economic Indicators (MEI) data base, from 1960 to the present. While prices or their logarithms are unanimously viewed as non-stationary, the issue is less clear for their year-to-year differences, in our case $\log(p_t/p_{t-4})$, which we will call 'inflation'. DF tests are not quite able to reject the null hypothesis of I(1),

while $\tilde{K}$ rejects stationarity for Austria but not for the other two cases. Therefore, German and French inflation point to the area of uncertain decision in the classical interpretation. In this regard, it is crucial that we use the significance points from the experimental Bayesian design. According to the tabulated values for the $\tilde{K}$ statistic, I(0) would be rejected for all countries.

Figure 10 shows that the criteria used in this paper indicate I(1) in all cases. French inflation comes close to the I(0) area if the comparison of p–values is adopted. We do not intend to give a final verdict on the properties of inflation. Anyway, one more differencing shifts all points far to the left and slightly downward, such that inflation is unlikely to be I(2).

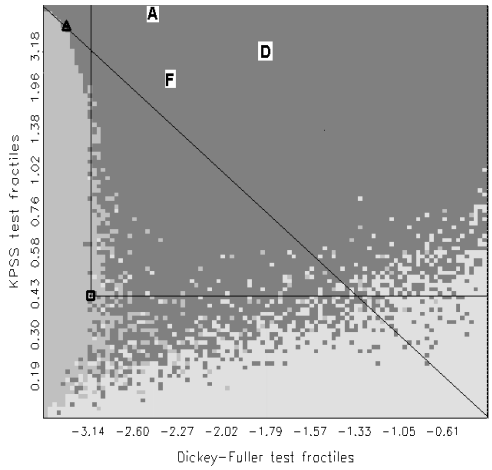

Figure 10: Austrian (A), German (D), and French (F) price inflation and its classification according to the p–value rule, joint confirmation, and Bayes tests, relying on a summary of the augmented DF and the $\tilde{K}$ statistics.

5 Summary and conclusion

In this paper, we compare approaches for obtaining a decision based on two univariate test statistics. These problems are common in statistical research, either because one wishes to consider two test statistics with similar properties but locally different power, or because one wishes to consider a very general maintained hypothesis, which is insufficiently represented by the null and alternative hypotheses that were used for constructing the test statistics. The latter case is of special interest, particularly if it involves an exchange of the roles of null and alternative hypotheses for the two test statistics.

Classical testing at a fixed significance level is the least satisfactory approach. While role reversal supports consistent decisions as $n \to \infty$, no advice is provided

in cases of conflicting outcomes. Two modifications were considered. If the decision is randomized in conflicting outcomes or, equivalently, some intermediate 'third decision' is adopted, the risk of incorrect classification is high. If the final decision rests on a comparison of p–values, this risk is more tolerable, even though the method is frowned upon by many theorists. However, in our simulations these p–values correspond to exact null distributions, which can be generated by computer-intensive procedures only, such as parametric bootstrapping. Using tabulated distributions would result in considerably higher Bayes risks.

Joint confirmation analysis provides an interesting solution to conflicting outcomes of component tests by adjusting the significance level to a point of one of the component alternatives by re-interpreting it as a null. Thereby, the approach succeeds in prescribing decisions for any outcome. However, because it remains within the boundaries of the frequentist viewpoint, we feel that it does not account fully for the information provided by the pair of observed statistics. Another disadvantage may be its inherent asymmetry. While the asymmetric treatment of null and alternative in the classical paradigm may reflect the need to put subject-matter theories to a test, asymmetry is difficult to support if two tests are conducted with role reversal of hypotheses. While the Bayes risk of joint confirmation exceeds the p–value method in some experiments, we note that this method rests on two tabulated values only, without any additional computation. Given the simplicity of the rule, joint confirmation is doing impressively well.

Bayes testing, with maps coded in the fractile space, succeeds in reaching fully consistent decisions and in automatically processing the provided information. Its drawback is its sensitivity to weight functions and its time-consuming simulation and evaluation. A major step in its widespread applicability could be the general standardization of weighting priors. Such standardization could provide a counterpart to the traditional classical significance levels. By construction, the Bayes test dominates in all experiments. Furthermore, analytical approximations of the Bayes-test decision contours may serve to explore optimal functions of the two test statistics. All other considered methods remain in the set of simple functions of the statistics τ_1 and τ_2, such as $\max(\tau_1, \tau_2)$. The contour curve is an implicit function $c(\tau_1, \tau_2)$, which suggests a new comprehensive test statistic that may serve as the basis of a univariate but risk-minimizing test. This is a possible direction for further research.

The Bayes-test approach is also the most flexible one if it comes to extensions of the maintained hypotheses. Instead of crudely viewing cases outside of the maintained hypotheses in the construction stage of the utilized test statistics as belonging 'rather' to the null or alternative, one may simply re-do the simulations on extended parameter spaces or add a third hypothesis to the decision set. The approach remains valid for any finite set of decision alternatives, much beyond the traditional setup of 'null' and 'alternative'. An application to a set of three alternatives was attempted by Kunst (2003). It is also conceivable to extend the approach in order to cover more than two univariate statistics, although the

graphical interpretation would then be lost.

The particular application of the principles—discriminating stationarity from difference stationarity—was selected, as it was intensely treated in time-series econometrics and constitutes one of the few examples for an application of the joint-confirmation approach in the literature. It should be noted that both the Bayes-test simulation design and the component tests can be improved. Recently, Leybourne *et al.* (2005) found that a relatively simple modification of the Dickey-Fuller test statistic due to Leybourne (1995) yields the most impressive power gains over several competing suggestions. The unusual null distribution of the Leybourne statistic is not a problem in the Bayes-test or in the joint-confirmation paradigm, as critical points or curves are determined by simulation. Such refinements are a topic for further research with this particular focus. An evaluation of the Bayes risk $g\left(\mathcal{L}_0\left(h_0, \Xi_1\right), \mathcal{L}_1\left(h_1, \Xi_0\right)\right)$ can demonstrate whether the gains in local power translate into global benefits for the underlying statistical decision problem. A similar remark holds with respect to all classification procedures that depend on single test statistics, such as the one by Stock (1994).

References

[1] Charemza, W.W., and E.M. Syczewska (1998) 'Joint application of the Dickey-Fuller and KPSS tests'. *Economics Letters* **61**, 17–21.

[2] Dickey, D.A., and W.A. Fuller (1979) 'Distribution of the estimators for autoregressive time series with a unit root'. *Journal of the American Statistical Association* **74**, 427–431.

[3] Dhrymes, P. (1998) *Time Series, Unit Roots, and Cointegration.* Academic Press.

[4] Ferguson, T.S. (1967) *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press.

[5] Hatanaka, M. (1996) *Time-Series-Based Econometrics: Unit Roots and Co-Integration.* Oxford University Press.

[6] Keblowski, P., and A. Welfe (2004) 'The ADF-KPSS test of the joint confirmation hypothesis of unit autoregressive root'. *Economics Letters* **85**, 257–263.

[7] Kunst, R.M. (2003) 'Decision maps for bivariate time series with potential threshold cointegration'. Working paper, Institute for Advanced Studies, Vienna.

[8] Kunst, R.M., and M. Reutter (2002) 'Decisions on seasonal unit roots'. *Journal of Statistical Computation and Simulation.* **72**, 403–418.

[9] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., and Y. Shin (1992) 'Testing the null hypothesis of stationarity against the alternative of a unit root,' *Journal of Econometrics* **54**, 159–178.

[10] Lehmann, E.L., and J.P. Romano (2005) *Testing Statistical Hypotheses.* Springer.

[11] Leybourne, S. (1995) 'Testing for unit roots using forward and reverse Dickey-Fuller regressions'. *Oxford Bulletin of Economics and Statistics* **57**, 559–571.

[12] Leybourne, S., Kim, T.H., and P. Newbold (2005) 'Examination of some more powerful modifications of the Dickey-Fuller test'. *Journal of Time Series Analysis* **26**, 355–370.

[13] Maddala, G. S., and I.M. Kim (1998) *Unit Roots, Cointegration, and Structural Change.* Cambridge University Press.

[14] Mueller, U.K. (2004) 'The Impossibility of Consistent Discrimination between I(0) and I(1) Processes'. Working Paper, Princeton University.

[15] Pantula, S.G. (1989) 'Testing for unit roots in time series data'. *Econometric Theory* **5**, 256–271.

[16] Perron, P. (1989) 'The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis'. *Econometrica* **57**, 1361-1401.

[17] Pratt, J.W., Raiffa, H., and R. Schlaifer (1995) *Introduction to Statistical Decision Theory.* MIT Press.

[18] Saikkonen, P., and R. Luukkonen (1993) 'Testing for a Moving Average Unit Root in Autoregressive Integrated Moving Average Models'. *Journal of the American Statistical Association* **88**, 596–601.

[19] Stock, J.H. (1994) 'Deciding between I(1) and I(0)'. *Journal of Econometrics* **63**, 105–131.

[20] Tanaka, K. (1996) *Time Series Analysis.* Wiley.

Author: Robert M. Kunst

Title: Approaches for the Joint Evaluation of Hypothesis Tests: Classical Testing, Bayes Testing, and Joint Confirmation

Reihe Ökonomie / Economics Series 177

ISSN: 1605-7996

Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 59991-555 • http://www.ihs.ac.at

ISSN: 1605-7996